SIGNAL PROCESSING IV: Theories and Applications
J.L. Lacoume, A. Chehikian, N. Martin, and J. Malbos (Editors)
Elsevier Science Publishers B.V. (North-Holland)

A VQ-based direct autocorrelation matching method for robust LP analysis of noisy speech

K.K. Paliwal

Computer Systems and Communications Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay-400005, India

ABSTRACT: Conventional linear prediction (LP) analysis methods (autocorrelation, covariance, Burg) do not perform satisfactorily for speech distorted by the addition of white noise. A number of methods which use indirect autocorrelation matching (or, equation-error norm minimization) have been proposed in the literature for robust LP analysis of noisy speech. But satisfactory performance is yet to be obtaied. In the present paper, a new method is proposed for robust LP analysis of noisy speech. This method uses direct matching of autocorrelation coefficients and exploits the properties of human speech production process in the form of vector quantization codevectors of autocorrelation coefficients representing different vocal tract shapes. These codevectors are derived in a once and for all training phase from the clean speech. For noisy speech, the present method is found to give better performance than the autocorrelation method and the high-order Yule-Walker method. In addition, it guarantees the stability of the estimated all-pole system.

1. INTRODUCTION

Linear prediction (LP) analysis of speech [1,2] has been applied extensively over the past fifteen years in various speech processing applications such as speech coding, speech recognition, speech enhancement and speaker recognition. In LP analysis of speech, the speech signal is modelled as the output of an all-pole (or, autoregressive (AR)) system. The aim of LP analysis is to estimate the model parameters (also called LP coefficients) correctly from the speech signal. A number of methods (such as the autocorrelation, covariance and Burg methods) have been reported in the literature to estimate these model parameters [1-3]. These methods work satisfactorily for clean speech recorded in quiet environments. However, in practice, the speech application systems have to operate in environments where background noise level is relatively high. In these situations, the resulting speech signal is noisy; i.e., it is corrupted by the addition of white noise. For such type of noisy speech signals, the conventional methods of LP analysis do not perform satisfactorily [4-6]. This happens because the noisy speech signal does not follow the AR model. In fact, it follows an autoregressive moving-average (ARMA) model [6]. The aim of the present paper is to develop a robust LP analysis method which works on noisy speech and estimates the AR part of the ARMA model correctly.

It is well known that in the autocorrelation domain only the zeroth autocorrelation coefficient is affected by the addition of white noise, while the other autocorrelation coefficients remain unchanged [6]. Thus, if the use of those Yule-Walker equations which contain zeroth order autocorrelation coefficient can be avoided while computing the AR parameters, the problem of robust LP analysis for noisy speech can be solved. This fact has been exploited by many authors [7-10] to develop robust AR spectral estimation methods for noisy signals. In these methods high-order Yule-Walker equations are used to estimate the parameters of the p-th order AR model. Gersch [7] and Chan and Langford [8] have used an exact number of p high-order Yule-Walker equations (from i=p+1 to 2p) to compute p AR parameters. Recently, Cadzow [9] has proposed the use of an overdetermined set of q($>$p) high-order Yule-Walker equations (from i=p+1 to p+q) for this purpose. Paliwal [10] has made use of the information about the pitch period P estimated from the given speech signal and proposed an instrumental variable-type of method where an overdetermined set of q high-order Yule-Walker equations (from i=P+1 to P+q) are used to compute p AR parameters. In both Cadzow's high-order Yule-Walker (HOYW) method [9] and Paliwal's pitch-based instrumental variable method [10], each high-order autocorrelation coefficient is predicted in terms of p preceding autocorrelation

coefficients and the AR parameters are computed by minimizing the total-squared autocorrelation prediction error. In other words, the AR parameters are estimated by minimizing the matching error between the high-order (i>p) autocorrelation coefficients (computed from the noisy speech signal) and their predicted values.

Recently, Kaveh and Bruzzone [11] have shown that these estimators using high-order autocorrelation coefficients matching are statistically inefficient. The loss of efficiency is more for the wide-band signals and less for the narrow-band signals. This loss of efficiency can be reduced significantly if the matching of low- as well as high-order autocorrelation coefficients is done for AR parameter estimation. However, the use of low-order autocorrelation coefficients for matching has the problem that the prediction of these low-order autocorrelation coefficients in terms of the past coefficients involves the zeroth autocorrelation coefficient which is seriously affected by the addition of white noise to the speech signal. A solution to this problem is to compensate the zeroth autocorrelation coefficient for the additive white noise and then use the matching of low- as well as high-order autocorrelation coefficients for AR parameter estimation. Some robust LP analysis methods using this type of solution are recently reported in the literature [14-16].

Though these autocorrelation matching methods result in better performance than the conventional LP analysis methods, their performance is still not satisfactory [9,10,12-14]. This is the main problem with these methods. Another problem with these methods is that the estimated AR model is not guaranteed to be stable which is an important requirement for speech analysis-synthesis applications [2]. Though the HOYW method [9] and the instrumental variable method [10] can be made to give stable AR model by enforcing the constraint of Levinson's recursion, the resulting method [15] has still the problem of poor performance.

It is not very difficult to find the reason why these autocorrelation matching methods give poor performance. It is because these methods use indirect matching between the autocorrelation coefficients (which are poorly estimated from the noisy speech signal) and their predicted values. Since these predicted values are obtained from the poorly estimated p preceding autocorrelation coefficients, the resulting AR parameter estimates (obtained from the autocorrelation prediction error minimization) are poor. In signal processing literature, these types of methods are called the equation-error methods [16] and it is well known that the direct matching methods (or, the fitting-error norm minimization methods) can, in general, result in better performance than these methods [17-19].

In the present paper, we present a direct autocorrelation matching method of robust LP analysis of noisy speech. This method uses the concept of vector quantization [20] for deriving the LP parameter vectors for the different possible shapes of the human vocal tract. In order to do it, a large corpus of clean speech data representing a variety of speakers (male, female and children) is recorded in a quiet environment. The p autocorrelation coefficients are computed for individual frames of the clean speech signal and the Linde-Buzo-Gray (LBG) algorithm [21] is used to cluster the vectors in the p-dimensional space [22]. This results in a codebook of N codevectors (where N can be typically 1024). These codevecotrs, in a sense, represent most commonly occuring shapes of the vocal tract. In order to compute LP parameters, for a segment of noisy speech, its (p+q) autocorrelation coefficients are computed. This (p+q)-dimensional vector is compared with each of the N codevectors in the codebook using a distance measure (defined in the next section). The p-dimensional autocorrelation coefficient codevector which shows the least distance with the input (p+q)-dimensional vector is then used to compute the required LP coefficients for the input segment of noisy speech.

Since this VQ-based method uses direct matching of low- as well as high-order autocorrelation coefficients and it exploits the properties of speech production process in the form of VQ codevectors representing different vocal tract shapes, it is expected to perform better than the above-mentioned indirect matching methods. In addition, it might be noted that this method always guarantees the stability of the estimated AR system.

2. THE VQ-BASED DIRECT AUTOCORRELATION MATCHING METHOD

Let the segment of observed noisy speech be

$$y_m = x_m + w_m, \qquad m=1,2,..,M,$$

where M is the segment duration, $\{x_m\}$ the uncontaminated (clean) speech (which follows the p-th order AR model) and $\{w_m\}$ the zero-mean white noise. The aim here is to estimate p AR parameters (or, LP coefficients) $\{a_i, i=1,2,..,p\}$ from the observed noisy speech signal $\{y_m\}$.

The VQ-based direct autocorrelation matching method proposed in the present paper requires an initial training process (which is done once and for all) before it can start estimating the AR parameters from noisy speech. The training process requires a large corpus of clean speech data collected from a large number of speakers of all sex in a quiet environment. (This is done to ensure all possible shapes of the vocal tract to be represented in the codebook.) The clean speech signal is analysed frame-wise and biased estimates of p autocorrelation coefficients are obtained from the Hamming-windowed speech signal. These

autocorrelation coefficients are normalized by dividing each of them by the zeroth autocorrelation coefficient. The LBG algorithm is applied on these p-dimensional autocorrelation coefficient vectors using the Itakura distance measure [22] and a codebook of size N is obtained. The p autocorrelation coefficients of each of the N codevectors are used to compute p AR parameters using the following Yule-Walker equations [1]:

$$\sum_{k=1}^{p} a_k r_j(|i-k|) = -r_j(i), \quad i=1,2,..,p,$$

for j=1,2,..,N. Using these AR parameters, the q extrapolated autocorrelation coefficients are obtained for each of the N codevectors from the following high-order Yule-Walker equations [9]:

$$r_j(p+i) = -\sum_{k=1}^{p} a_k r_j(p+i-k), \quad i=1,2,..,q,$$

for j=1,2,..,N.

This completes the training process for the present method. At the end of training process, a lookup-table is generated having N entries, each entry having with it (p+q) normalized autocorrelation coefficients and p AR parameters.

In order to compute the AR parameters of the noisy speech signal $\{y_n\}$, the present method requires the following two steps. In the first step, the biased estimates of (p+q)-autocorrelation coefficients $\{R(i)\}$ are computed from the noisy speech signal. In the second step, this (p+q)-dimensional autocorrelation vector is compared with each of the N normalized autocorrelation vectors in the lookup-table (generated during the training phase) using the following distance measure:

$$d_j = \sum_{k=1}^{p+q} \{R(k) - c_j r_j(k)\}^2, \quad j=1,2,..,N,$$

where c_j is the scale factor whose optimum value is given by

$$c_j = [\sum_{k=1}^{p+q} R(k) r_j(k)] / [\sum_{k=1}^{p+q} \{r_j(k)\}^2].$$

Here, it might be noted that in this definition of distance measure use of the zeroth autocorrelation coefficient is deliberately avoided to take care of additive white noise. The address of the autocorrelation vector which gives the smallest distance is noted and the p AR parameters associated with this address are found from the lookup-table. These are then the required AR parameter estimates for the observed noisy speech signal $\{y_n\}$.

3. RESULTS

The VQ-based direct autocorrelation matching method is studied here on a highly limited speech data and at present only the preliminary results are available about the performance of this method. More detailed investigations are in progress. In order to put the VQ-based method in a proper perspective, its performance is compared with that of the autocorrelation method [1] and the high-order Yule-Walker (HOYW) method [9].

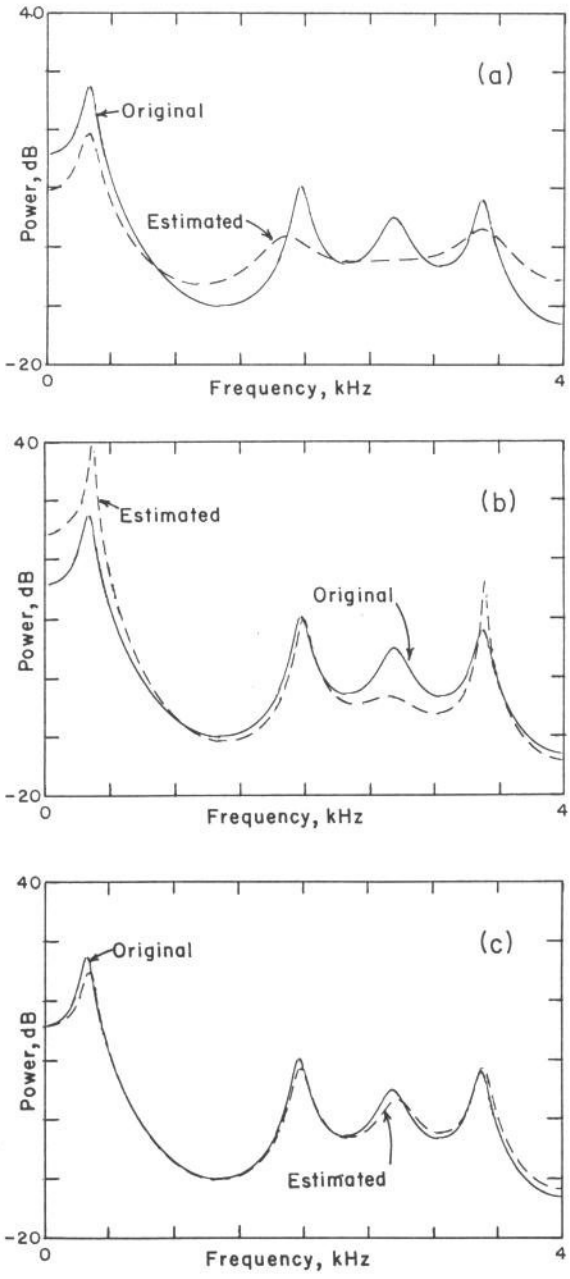

Fig. 1. Power spectrum estimates of a 20-ms segment of noisy speech with SNR=10 dB using a) the autocorrelation method, b) the HOYW method and c) the VQ-based method.

In the training phase of the VQ-based method, 40 sec of clean speech from a single male speaker is recorded in a noise-free room. Clean speech is digitized at 8 kHz sampling rate. Digitized speech is analysed at the rate of 100 frames per second with frame duration of 20 ms. A 10-th order LP analysis is performed. Using the LBG algorithm, a codebook of N=256 codevectors is generated. For testing the performance of the VQ-based method, a 20 ms segment of clean speech from vowel /i/ is taken and corrupted by the addition of white Gaussian noise making its signal-to-noise ratio (SNR) equal to 10 dB.

Resuts are shown in Fig. 1. It can be seen from this figure that the autocorrelation method gives very poor performance in the sense that higher formants are not visible in its spectral estimate. The HOYW method improves the situation, but its spectral estimation performance is still poor. The VQ-based method results in power spectrum estimate much better than the autocorrelation and the HOYW methods.

4. CONCLUSIONS

In this paper, a VQ-based direct autocorrelation matching method is proposed for robust LP analysis of noisy speech. This method exploits the properties of human speech production process and ensures the stability of the estimated all-pole filter. In terms of spectral estimation performance, this method is found to be better than the conventional autocorrelation method and the HOYW method.

It might be noted here that these conclusions are based on preliminary results derived from a limited speech data of a single speaker. More detailed study is required for putting this method on a sound foundation. For example, it is necessary to study the method in a speaker independent mode and see the effect of codebook size on its performance. It might also be noted that the method uses the Itakura distance measure during the training phase and the long autocorrelation distance measure in the actual operation. The autocorrelation distance measure has certain limitations. Investigation of more effective alternate distance measures [23] is necessary to improve the performance of the present method further.

REFERENCES

[1] J. Makhoul, Proc. IEEE, 63, 1975, pp. 561-580.

[2] J.D. Markel and A.H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, Berlin, 1976.

[3] K.K. Paliwal and P.V.S. Rao, Signal Processing, 4, 1982, pp. 59-63.

[4] M.R. Sambur and N.S. Jayant, IEEE Trans. ASSP-24, 1976, pp. 488-494.

[5] J. Tierney, IEEE Trans. ASSP-28, 1980, pp. 389-397.

[6] S.M. Kay, IEEE Trans. ASSP-27, 1979, pp. 478-485.

[7] W. Gersch, IEEE Trans. AC-15, 1970, pp. 583-588.

[8] Y.T. Chan and R.P. Langford, IEEE Trans. ASSP-30, 1982, pp. 689-698.

[9] J.A. Cadzow, Proc. IEEE, 20, 1982, pp. 907- 939.

[10] K.K. Paliwal, Proc. EUSIPCO, Hague, Sept. 1986, pp. 593-596.

[11] S.P. Bruzzone and M. Kaveh, IEEE Trans. ASSP-32, 1984, pp. 701-715.

[12] V.K. Jain and B.S. Atal, Proc. ICASSP, Tampa, Mar. 1985, pp. 473-476.

[13] K.K. Paliwal, Proc. ICASSP, Tokyo, Apr. 1986, pp. 1369-1372.

[14] K.K. Paliwal, Digital Signal Processing-87, V. Cappellini et al. (Eds.), North-Holland, Sept. 1987, pp. 739-743.

[15] K.K. Paliwal, Proc. EUSIPCO, Hague, Sept. 1986, pp. 295-298.

[16] L.B. Jackson, Digital Filters and Signal Processing, Kluwer Academic, Boston, 1986 (Chapter 10).

[17] K. Steiglitz and L.E. McBride, IEEE Trans. AC-10, 1965, pp. 461-464.

[18] A.G. Evans and R. Fischl, IEEE Trans. AU-21, 1973, pp. 61-65.

[19] R. Kumaresan et al., IEEE Trans. ASSP-34, 1986, pp. 637-640.

[20] J. Makhoul et al., Proc. IEEE, 73, 1985, pp. 1551-1588.

[21] Y. Linde et al., IEEE Trans. COM-28, 1980, pp. 84-95.

[22] A. Buzo et al., IEEE Trans. ASSP-28, 1980, pp. 562-574.

[23] Y. Ephraim et al., Proc. ICASSP, Dallas, Apr. 1987, pp. 1324-1327.